# Ecommerce Capstone Project

## Background - Business Understanding

ElectroKart is an e-commerce firm specialising in electronic products. Over the last one year, they had spent significant amount of money in marketing. Occasionally, they had also offered big ticket promotions (similar to the Big Billion Days). They are about to create a marketing budget for the next year which includes spend on commercials, online campaigns, and pricing and promotion strategies. The CFO feels that the money spent over last 12 months on marketing was not sufficiently impactful and that they can either cut on the budget or reallocate it optimally to improve the revenue response.

Imagine that you are a part of the marketing team working on budget optimization. You need to develop a market mix model to observe the actual impact of different marketing variables over the last year. Using your understanding of the model, you have to recommend the optimal budget allocation for next year.

**Data Understanding**

**You can download the dataset from this** [**link.**](https://drive.google.com/drive/folders/0B22sHi5RF-VLS1paeTJGWURReTg?usp=sharing)

You have to use the data are from July 2015 to June 2016. The data consists of the following types of information:

**Order level data**

* FSN ID: The unique identification of each SKU
* Order Date: Date on which the order was placed
* Order ID: The unique identification number of each order
* Order item ID: Suppose you order 2 different products under the same order, it generates 2 different order Item IDs under the same order ID; orders are tracked by the Order Item ID.
* GMV: Gross Merchandise Value or Revenue
* Units: Number of units of the specific product sold
* Order payment type: How the order was paid – prepaid or cash on delivery
* SLA: Number of days it typically takes to deliver the product
* Cust id: Unique identification of a customer
* Product MRP: Maximum retail price of the product
* Product procurement SLA: Time typically taken to procure the product

Apart from this, the following information is also available:

* Monthly spend under various advertising channels
* Days when there was any special sale
* Monthly NPS score – this may work as a proxy to ‘voice of customer’

You can download the product details from the [link](https://docs.google.com/a/ueducation.in/document/d/1v-UHk_96xg06vLTPZpB0UzZH7tm8eVUR-ITPVu1wwb0/edit?usp=sharing) below.

## Data Preparation

You have to create market mix models for three product sub-categories - **camera accessory, home audio and gaming accessory.** Also, the models have to be built at a **weekly level**.

Create **3 different subsets** of the data for each of these product sub-categories. All further analysis and modeling needs to be done at the **product sub-category level.**

Create all the relevant KPIs and the derived KPIs you can think of, for example lags, adstocks, promotion etc.

# A suggestive list of engineered KPIs that you should create:

1. Promotion offered = (MRP - list price)/MRP where list price = GMV/units sold
2. List price inflation: wrt previous week, wrt average price of previous 2/3/4/5/6 weeks, wrt moving average of previous week, wrt moving average of previous month.
3. Lag sale, i.e. the past value of GMV (will be useful for Koyck and distributed lag models)
4. Lag promotion, i.e. the past value of promotion
5. Weekly percentage change in promotion, monthly percentage change in promotion etc.
6. Adstock of each of the ‘commercial spends’ (Since we don’t have TRP/ Impression information, we will assume spend is directly proportional to the impression)
7. Log of **adstock,** log of other independent variables (log adstocks, log NPS etc.)
8. To be more creative – you may create a product level variable called product premium-ness – based on the MRP and units sold under each product vertical, you may tag whether the product is a **mass market**, **aspiring** or **premium** product
9. If you can source the total revenue from the electronic market in India (Month by month) from the internet, then you may include that too as a macroeconomic variable
10. Daily **temperature** and **rainfall** – would help you to proxy weather and seasonality effects
11. You may as well get an Indian holiday calendar for this period and include that as a variable

**Modelling**

You are expected to build five kinds of models:

1. Basic linear model where GMV is the dependent variable and other base level KPIs are independent
2. Multiplicative model where all dependent & independent variables go in Ln (Natural log transformation) form
3. Koyck model
4. Distributed lag model
5. A combination of distributed lag and multiplicative model

**Presentation of Results**

After building the models, you have to choose the best ones for each of the three product sub-categories and explain them to the CMO, CFO etc.

Create a powerpoint presentation explaining the best models. The audience should be able to intuitively understand the variables and their importance in predicting future sales. Point out any surprising or unexpected trends you notice.

### Mentorship Sessions

Given below is the schedule for the mentorship sessions and what milestone you are expected to achieve before the sessions.

### Mentorship Session

### Agenda

**Call 1**

Understanding of the business problem and data. Overall structure to solve the problem.

**Call 2**

Engineering the KPIs and results of the initial linear model.

**Call 3**

Results of the models (variables, KPIs and validation results), suggestions to improve the models. The final choice of model to be deployed.

**Call 4**

Presentation of results to the business - important variables, results (in simple terms) etc.

**Do's and Don'ts for Mentorship Sessions**

* Be prepared for the session as per the agenda prescribed above
* You would need to drive these sessions and extract the most out of these sessions.
* Ask relevant and worthwhile queries that help you in clarifying the doubts
* Do not push the mentors to provide solution
* Do not ask the mentors to review your R-codes
* Respect the time of mentors - if any mentorship session goes unattended by all the members of a group, no further mentor support will be available to the group

### Rubrics

|  |  |  |
| --- | --- | --- |
| **Stage** | **Meets expectations** | **Does not meet expectations** |
| Data understanding, preparation and EDA | All data quality checks are performed and data preparations steps are performed correctly.  Relevant EDA is done using plots and summaries. The insights from EDA should be clearly derived and explained.  All the relevant KPIs and derived KPIs have been engineered. | Data analysis is not done at weekly level or the data from different tables are not analysed at the same base level. |
| Model building | The model building process, i.e. variable selection, tuning and model evaluation steps are correct and in accordance with model selection principles.  Model evaluation is conducted correctly and tests the accuracy, stability and generalisability of the model.  Satisfactory R square value is achieved (at par with the best possible on this dataset).  The final model is chosen based on relevant metrics and the justification for the same is provided.  Model and the important variables are interpreted correctly (from a business perspective).  All 5 types of models are built for each of the 3 sub categories. | All the required models are not built.  Appropriate steps for variable selection, tuning and evaluation are not followed.  The final model’s evaluation metrics are sub-optimal.  The final variables are not interpreted correctly.  All types of models are not built or some of them are incorrectly built. |
| Presentation of results | The presentation has a clear structure, is not too long, and explains the most important results concisely. The final recommendations include a brief explanation of the important variables, the model, the financial implications of using the model.  Overall, the presentation proves to be impactful to the business. | Results are presented without appropriate visualisations, explanation or learnings from the same.  The presentation does not clearly show the positive impact of the model to the business. |